Creation and Maintenance of Helix, a Web Based Database of Medical Genetics Laboratories, to Serve the Needs of the Genetics Community

Peter Tarczy-Hornoch MD, Maxine L Covington, Joseph Edwards PhD,
Paul Shannon, Sherrilynne Fuller PhD, Roberta A Pagon MD.
Division of Biomedical Informatics and Division of Genetics
University of Washington and Children's Hospital Regional Medical Center, Seattle WA.

Helix (healthlinks.washington.edu/helix) is a web accessible database that serves as the main U.S. directory of laboratories offering genetic testing. The database was designed to address the previously unmet need for a centralized, continuously updated source of information about clinical and research genetic testing to keep pace with the rapid rate of gene discovery resulting from the Human Genome Project. The Helix project began in 1992 at the University of Washington and Children's Hospital and Regional Medical Center. It has evolved from a single user stand alone relational database to a fully Web enabled database queried and maintained via the web and linked to other web accessible genomic databases. As of February, 1998 it lists more than 500 diseases and 290 laboratories, with over 5,200 registered users making ~250 queries/day (90% via the Internet). We describe the iterative design, implementation, population and assessment of the database over a six year period.

INTRODUCTION

The Helix Genetic Testing Resource is a database designed to meet the growing need[1] of a broad community of users for information on availability of genetic testing. New gene discoveries resulting from the Human Genome Project[2] have led to advances in genetic testing that can improve medical care and expand personal choices for individuals with inherited disorders[3]. Genetic testing is becoming increasingly relevant to clinicians as a cost-effective, sensitive, non-invasive diagnostic tool, and to patients as the primary method for presymptomatic diagnosis, carrier detection, and prenatal diagnosis of genetic diseases[4]. The use of genetic testing is complicated by: (1) rapid transition from the research laboratory to clinical practice, which raises concerns about test validity and laboratory proficiency[5]; (2) the number and diversity of laboratories offering testing, which raises concerns about the ability to identify and access laboratories testing for rare disorders[1]. The molecular biology research community widely uses on-line databases[6,7,8,9]; however, these are not suited to the clinical genetics community. Prior to Helix there was no centralized listing of available genetic testing; the only on-line clinical genetic resource was a diachronic catalog of clinical phenotypes and genes, OMIM (On-Line Mendelian Inheritance in Man)[10].

EVOLVING REQUIREMENTS

The initial implementation of Helix (Version 1.0 11/92-10/96) as a stand-alone single user relational database (in Foxpro) fulfilled the initial requirement for a "yellow pages" listing genetic diseases and the laboratories testing for them. The targeted audience of genetics healthcare providers and researchers placed requests by phone or fax and received a faxed report within one working day. All database maintenance (getting information from labs/users) and queries were handled by one full time staff member. Based on a user survey in Dec 1994 of the 1920 registered Helix users (20% response rate) and 197 laboratory directors (32% response rate), 90% of respondents documented the significant utility and 92% the uniqueness of the database and, for 77% of labs, increased referrals. The survey also identified areas for improvement, most importantly the need for direct user access to the database 24 hours/day 7 days a week (particularly for clinics offering prenatal diagnosis).

The second implementation of Helix as a Web searchable database (Version 2.0 10/96-12/97) addressed the need for 24x7 access. This interim implementation involved a biweekly mirroring of the existing Foxpro database content to an Informix database with Web access via a CGI interface to a C/ESQL interface to the database. Based on requests by laboratory directors to minimize direct patient inquiries and to assure that patients would receive information about testing in a clinical setting, Helix Web access was restricted to registered users via a combination of htaccess and user identifiers. Separate Informix (Web) and Foxpro (assisted) usage logs were maintained by the databases. The site also provided access to on-line user/laboratory registration and feedback forms, and a link to a limited directory of regional clinical contacts.

Helix 2.0 had unresolved administrative needs: (A1) multi-administrator capability, necessitated by growth of content and usage and the addition of a second Helix staff member; (A2) better ways to maintain laboratory/test contact information; (A3) more elaborate and varied reports of Helix usage and content. Unresolved content needs included: (B1) documentation of clinical laboratory certification; (B2) more detail on testing methodologies; (B3) links

to citations; (B4) links to laboratory web sites and e-mail addresses; (B5) links to molecular databases. The major unresolved clinical need (B6) was to relate to gene-based (genotype) information to clinically-based (phenotype) information.

METHODS & IMPLEMENTATION

The current implementation of Helix (Version 3.0 12/97-present) addressed the limitations of the first two versions by using a new data model and a web based client/server architecture.

A1: Distributed multi-user database administration was accomplished by using a large sophisticated Java applet that connects to the database server via a custom C/ESQL connection server developed by the Helix Informatics Team. User queries rely on C/ESQL scripts with CGI interfaces. The database running on an HP J282 server consumes more than 50% of server capacity at peak load. Security for data entry is provided by a Java/database implemented user id/password as well as IP and domain restriction.

A2: The problems of duplicate, at times inconsistent, information was addressed by normalizing the database design. All information that pertains to a laboratory and all the test packages it offers are now attributes of the *Laboratory* rather than of the *Test Package* (Figure 1). Contact information is now stored in a *Rolodex* entity which in general is linked to the *Laboratory* since it is the same for all *Test Packages* offered by a laboratory. Provision is also made to permit a *Test Package* to have its own contact information when necessary (including its own director, contact and specimen contact).
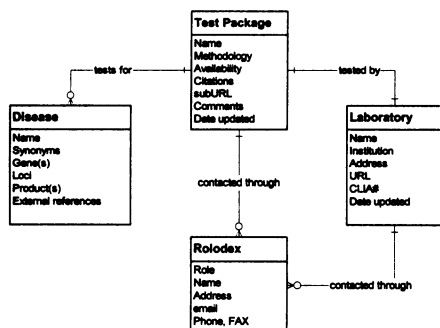

Figure 1: Key entities in current schema

The 3.0 schema avoids duplicating information about people and addresses (Figure 2). It permits a person to have an unlimited number of roles with shared or unique contact information as appropriate (e.g. to be the director of one laboratory while being the contact only for a single test package offered by that laboratory and to have different addresses for these two roles). Data maintenance tools were written to permit the Helix staff to identify all roles held by a person, all contact information for a person, all uses of a given address, etc. and to globally update and

cross reference information when necessary. If the manager tries to delete a record that is still being referenced, a pop-up window indicates what records are still using the reference and it can not be deleted.
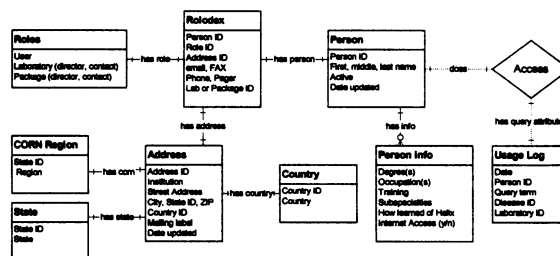

Figure 2: People, roles, addresses

A3: The three major functions of the reports are: (1) Improved database maintenance: For example, the Comprehensive Lab Report, a printout of the complete Helix data on an individual laboratory, allows the laboratory director to review his/her entries in aggregate rather than a single *Test Package* at a time. (2) Summary reports on genetic testing availability: Unique Helix data can be sorted and displayed to answer such questions as: How many laboratories perform clinical genetic testing? For how many (and which) genetic diseases does a clinically useful genetic test exist? Of these, how many are diagnostic (direct DNA analysis) and how many are for carrier detection (linkage analysis)? (3) Indicators of genetic testing utilization patterns: Helix use statistics directly answer such questions as: Who is using Helix to access information on genetic testing? How many physicians, who are not geneticists, are registered to use Helix? What are their specialties? How often do they use Helix? The answers to these types of questions indirectly suggest patterns of utilization of genetic testing in the broader medical community.

B1: The distinction between clinical and research availability is made based on the HCFA CLIA regulations[11]. Clinical laboratories (those providing diagnostic testing for patients) are required to provide a CLIA number (Figure 1).

B2: A table driven, expandable set of *Test Package* attributes now captures detailed specific information on testing methodology (Figure 1). A given laboratory's *Test Package* for a particular disease can have 1..n methodologies which include: DNA analysis (direct, linkage, methylation, uniparental disomy, X inactivation, trinucleotide repeat); molecular cytogenetics (FISH); biochemical (analyte); and protein analysis (enzyme assay).

B3/B4: Literature citations and links to PubMed[12] are provided. Mailto: links are provided to all e-mail addresses for contacts (Figure 2) and multi-level links are provided to laboratory web sites (Figure 1).

342

B5/B6: The disease entity (phenotype) was expanded to include disease-specific links and interfaces to other on-line clinical genetic databases[10,13]. Data structures for genotype information (Figure 3) (*Gene*-gene name, *Locus*-chromosomal locus, *Product*-protein or other product) were created and linked to phenotype (designated diseases). This genotype information provides a way (albeit limited) to express more specific disease (phenotype) and test (genotype) relationships and another means for searching Helix. We have implemented a general purpose mechanism to permit each of these entities to link to external references[6,7,8,9,10,12,13] with provisions to support generic http interfaces as well as API-type interfaces.
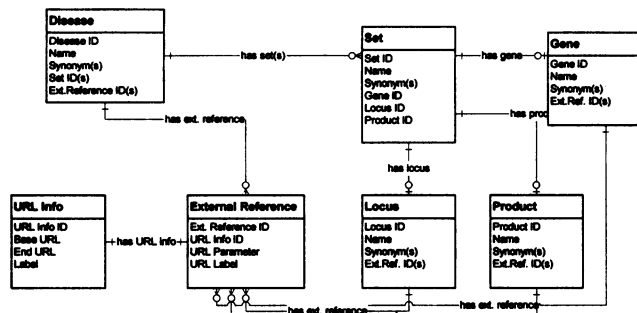

Figure 3: Genotype entities

This expanded genotype/phenotype linkage is important since there are times when the relationship of phenotype to genotype is more complex than the classic one gene/one product/one disease model. In some cases different mutations of the same gene can cause multiple diseases. In other cases, the same disease entity can be caused by different genes. For example, the disease tuberous sclerosis is caused by mutations in the genes TSC1 [locus 9q34, product unknown], TSC2 [locus 16p13, product tuberin], and TSC3 [locus 12q22-24, product unknown]. The genotype (TSC1, TSC2 or TSC3) may then be specified for each laboratory providing testing for tuberous sclerosis, permitting clinicians to select laboratories offering the most suitable gene-specific testing.

Conversion of the data from version 2.0 to 3.0 presented some challenges. For example, in the 2.0 schema, user information was stored in one table that contains the user's name, address, contact information, degree information, title, etc. In the 3.0 schema, user information is stored in the rolodex and person_info tables. Addresses, states, countries, degrees, and occupations are stored in lookup tables and referenced by id numbers in those tables. In order to convert the data, countries, states, degrees and occupations had to be extracted, given unique id numbers and stored in lookup tables, and the corresponding user data fields had to be filled in with the correct id number. The addresses were extracted

and state and country fields were changed to an id number. The corresponding address_id was then associated with a user record. The resultant data were written to SQL scripts that were used to load the data into the new tables using dbaccess. Perl scripts were written to automate this process and were used to periodically update the 3.0 database while it was being tested. The contact information from tests presented additional challenges since the 2.0 database integrity had been compromised. To solve this problem we used a combination of intelligent pattern matching software and manual review of the contact mapping by Helix staff.

RESULTS

A. Usage

Helix content and usage has grown steadily (Table 1) reflecting it unique role in the genetics community[14]. Although 90% of the use is via the Internet, phone/fax mediated searches continue to serve a significant number of users.

Date	7/93	10/96	12/97	2/98
Users	800	3,700	5,000	5,200
Diseases	111	420	485	520
Labs	110	250	300	290
Monthly Searches	~200	~540	~2860	~5000
% Internet searches	0%	0%	90%	90%

Table 1: Helix content and usage over time

Genetics healthcare providers use Helix to identify laboratories to serve their patients' medical needs. Laboratory directors list their laboratories in the Helix directory, use Helix as a referral source when they cannot provide a service themselves, and identify other laboratories performing similar testing for informal quality assurance purposes. For researchers, the benefit to listing in Helix is increased ascertainment of families through high visibility at the point of care and the identification of potential collaborators. Helix maintains high visibility among genetics professionals through staff monitoring of the National Society of Genetic Counselors listserv, and through an exhibitor's booth at the annual American Society for Human Genetics Meeting. Visibility to the broad medical community has occurred through links from relevant web sites and by citations in the medical[1,14,15] and lay[16] literature.

The success rate for Helix staff-mediated searches is 81% (that is, 81% of the time the requested disease is in fact listed in Helix), whereas only 47% of searches via the Internet find a disease. The reason for this discrepancy is under investigation (detailed human and computer analysis of logs of search terms/results). It may reflect the changing demographics of Helix users. Initially 100% of Helix users were genetics professionals, whereas in 2/98

over 20% are not genetics professionals. In Helix 3.0, the audiences include: (1) non-geneticist healthcare providers (driven by the growing realization of the role of genetics in common diseases[1,15,17]); (2) healthcare policy makers; and (3) the public (also driven by the growing role of genetics), for whom an appropriate limited view of Helix is being developed.

B. Reports

Two summary reports of genetic testing availability have been prepared for inclusion in chapters (in preparation) on genetic testing in Current Protocols in Human Genetics (JC Wiley) and Scientific American Medicine. The use of summary reports for healthcare policy makers and test reimbursement policy development is being explored.

C. Evaluation

In January 1998, the 1,700 Helix users with e-mail addresses on file were invited to submit letters commenting on the value of Helix. 15% responded (Table 2) all expressing enthusiastic support for Helix. The demographics and usage rates of respondents and non-respondents were similar.

FREQENCY OF USE	OCCUPATION	AFFILIATION
40% Frequent, routine	33% Genetic counselors	52% Universities
21% Daily	30% MD geneticists	24% Private practice
20% Daily-weekly	15% PhD researchers	11% HMOs
14% Weekly-monthly	8% MD other	10% Labs
5% Occasional	8% Helix lab directors	3% Government
	6% Other	

Table 2: Demographics (by % respondents)

NEED	BENEFITS
77 Rapid proliferation of genetic information	111 Improved patient care
	70 Furthering genetic research
56 Complexity and rarity of genetic diseases	39 Timeliness for prenatal diagnosis
29 Cumbersome methods used before Helix	19 Accessibility in rural/isolated areas
21 Wide dispersal of laboratories	♦
10 Standard for trainees	
5 Often noted on NSGC listserv	

Table 3: Summary of comments (# of respondents)

Respondents spontaneously described in very similar terms the unique problems in genetics ("need") which make Helix so useful ("benefits") (Table 3).

DISCUSSION

Helix, a national database of genetic testing laboratories, is an important and heavily used resource that is well suited to the web environment as demonstrated by the continued growth of Helix and in particular the very heavy usage of the Internet accessible version (Table 1). In 1996 the only other analog (EDDNAL – European Directory of DNA Laboratories[18]) was independently created. EDDNAL differs in that at present it provides limited contact information with no methodology or availability information, no genomic information, no links to other on-line molecular or clinical genomic databases, and no user-specific usage logs.

There are pros and cons to our implementation of Java based database maintenance tools. We wrote JDBC compliant methods, thus our custom JDBC driver/broker is very modular and we could easily point to a different database. The code necessary to develop a full fledged data entry/database maintenance application is complex enough that it pushes the limits of current browser Java Virtual Machine implementations - the current version runs optimally on a Pentium 166 system with 32M of RAM with a direct Internet connection. New generations of Java VMs and JIT compilers will help address this. The advantages include: 1) downloading the application from the server to the client when needed without installation on the client PC; 2) immediate availability to users of updates; 3) the use of live client side forms permitting data validation prior to submission and the use of live pick lists. Disadvantages of the Java solution include: 1) insufficiently robust commercial JDBC driver/brokers for utilization on the development platform (database transactions were are not always successful resulting in loss of data and corruption of database integrity); 2) need for custom development of a database connection and result classes; 3) inability of some Java enabled browsers (e.g. on the Macintosh) to support a socket connection resulting in 4) lack of true platform independence, requiring programmer intensive software solutions. The maturation of Java and object databases (in particular JDK 1.2 persistent classes) presents a probable solution to many of these issues.

LIMITATIONS AND PLANS

We plan two fundamental enhancements to the Helix database: 1) to extend the genotype/phenotype model in order to 2) link the information in the Helix database on test availability to information on applicability of genetic testing.

The need to extend the genotype/phenotype model stems from our observation that the causal relationship between genotype and phenotype is fluid and continually refined by clinical and molecular genetic research. Clinical diagnosis and molecular genetic testing represent quite different approaches to medical practice and develop with their own techniques, pace, and ontologies. Our new data model supports both domains independently via an elaboration of our existing schema, and supports the connection between the two with "causality maps". The evidence suggests that over time most genetic diseases (clinical phenotypes) will be ever more precisely mapped to disease-causing alleles (genotypes). Our new prototype supports this. Conversely, it also supports the aggregation (or lumping) of phenotypes and of genotypes, and allows imprecise or unknown causality, since these usually precede the appearance of a final causal model.

Helix needs to be linked to information on the proper application of genetic testing. While molecular genetic testing is a powerful clinical tool, it is apparent that some clinicians have difficulty applying molecular genetic testing to patient care[17]. Specific problems noted were: utilization of inappropriate testing strategies (21%), inability to interpret test results (32%) and failure to provide genetic counseling (81%). To address the need for quality information relating new advances in genetic testing to the diagnosis, management and counseling of patients, we have developed (currently in advanced prototype form) the expert-authored, peer-reviewed Genline Medical Genetics Knowledge Base, which also fulfills the standards for quality medical information[19]. This on-line database is attempting to meet the challenge that has been issued to replace our current dinosaurs[20] with new information tools that are electronic, portable, fast, easy to use, linked to medical knowledge bases[21], and of use to patients as well as clinicians. Through integration with Genline, Helix will meet the needs of clinicians, researchers, policy makers and the public for information on the availability and applicability of genetic testing.

CONCLUSIONS

The success of Helix demonstrates that the web is an effective way to deliver complex, rapidly changing information to the medical community. It further demonstrates the feasibility of using Java for group distributed platform independent administration and maintenance of a complex database. Porting the database to the Internet necessitated changes to the underlying data model to store information about interfacing to entries in other Internet genomic databases. Though Java and JDBC offered the promise of quickly implementing a seamless platform independent data entry to our experience was that the technology was immature and custom solutions were thus necessary. Persistent classes in JDK 1.2 potentially will resolve many of these issues.

The Helix database structure needs to be flexible to permit it to accommodate our evolving understanding of the genetic basis of disease as well as advances in genetic testing technologies. Integration of genetic testing (Helix) and application (Genline) is essential to good patient care and will necessitate the development of a shared data model.

References

1. Cotton P. Prognosis, diagnosis, or who knows? Time to learn what gene tests mean. *J Amer Med Assn* 1995;273: 93-95
2. Watson JD. The Human Genome Project: Past, present and future. *Science* 1990;248:44-49
3. Fink L, Collins FS. The Human Genome Project: View from the National Institutes of Health. *J Am Med Womens Assoc* 1997;52:4-7
4. Bird TD, Bennett RL. Why do DNA testing? Practical and ethical implications of new neurogenetics tests. *Ann Neurol* 1995;38:141-146
5. Holtzman NA, Watson MS: Promoting safe and effective genetic testing in the U.S. NIH, 9/1997.
6. Benson DA, Boguski MS, Lipman DJ, Ostell J. Genbank. *Nucleic Acids Res* 1997;25;1-6
7. Fasman KH, Letovsky SI, Li P, Cottingham RW, Kingsbury DT. The GDB Human Genome Database 1997. *Nucleic Acids Res* 1997;25: 72-81
8. Stampf DR, Felder CE, Sussman JL. PDBbrowse-- a graphics interface to the Brookhaven Protein Data Bank. *Nature* 1995;374:572-574
9. Appel RD, Bairoch A, Hochstrasser DA. A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Bioche Sci* 1994;19:258-260
10. Pearson P, Francomano C, Foster P, Bocchini C, Li P, McKusick V. The status of Online Mendelian Inheritance in Man (OMIM) medio 1994. *Nucleic Acids Res* 1994;22:3470-3473
11. Federal Register 1992;57:40
12. PubMed. www.ncbi.nlm.nih.gov/PubMed/
13. Genline. healthlinks.washington.edu/genline
14. Sikorski R, Peters R. Genomic medicine: Internet resources for medical genetics. *JAMA* 1997;278:1212-13
15. Bird TD, Bennett RL. Why do DNA testing? Practical and ethical implications of new neurogenetic tests. *Ann Neurol* 1995;38:141-146
16. Beardsley T. Vital data: trends in human genetics. *Scientific American* 1996;March:100-105
17. Giardiello FM, Brensinger JD, Petersen M, et al. The use and interpretation of commercial APC gene testing for familial adenomatous polyposis, *N Engl J Med* 1997;336:823-827
18. EDDNAL. www.eddnal.com
19. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet. *JAMA* 1997;277:1244-1245
20. Weatherall DJ, Ledingham JGG, Warrell DA. On dinosaurs and medical textbooks. *Lancet* 1995;346:4-5
21. Smith R. What clinical information do doctors need? *BMJ 1996;*313:1062-1068